2008

Maximum Likelihood Logistic Regression With Auxiliary Information

R. Chambers
University of Wollongong, ray@uow.edu.au

S. Wang
Texas A&M University

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

12-08

Maximum Likelihood Logistic Regression With Auxiliary Information

Ray Chambers and Suojin Wang

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Maximum Likelihood Logistic Regression With Auxiliary Information

Ray Chambers
Centre for Statistical and Survey Methodology, University of Wollongong
Wollongong NSW 2522, Australia
ray@uow.edu.au
and
Suojin Wang
Department of Statistics, Texas A&M University
College Station, Texas 77843, U.S.A.
sjwang@stat.tamu.edu

Summary

In this paper we use the general approach to maximum likelihood estimation for complex survey data described in Breckling *et. al.* (1994) to develop methods for efficiently incorporating external population information into linear logistic regression models fitted via sample survey data. In particular, we use innovative saddlepoint and smearing methods to derive highly accurate approximations to the score and information functions defined by the model parameters under random sampling and under case-control sampling when auxiliary data on population moments are available. Simulation-based results illustrating the resulting gains in efficiency are provided.

Some key words: Sample survey; Case-control; Saddlepoint approximation; Smearing estimates; Score function; Information function.

1. Introduction

Analysis of survey data does not happen in a vacuum. A model for the probability that a woman uses a particular contraceptive method will typically depend on a number of factors, e.g. her age, her education level, her labour force status, her household income, her ethnic background and her access to family planning information, perhaps measured by presence or

absence of a family planning clinic within a specified distance of her home. All of these variables are measured for women taking part in the survey, and the classical approach is to consider them 'in isolation' in the modelling process, implicitly assuming that the model fitted to these sample data is also appropriate for the population from which the sample is drawn. Sometimes, if this is felt to be too big an assumption, and survey weights are available, these are included in the model fitting process, assuming that they correct the parameter estimation process for potential sample selection bias.

However, we typically know a lot more about the target population than just the data observed in the survey. In particular we may know the total number of women in the population, the proportion of women in the population that use the contraceptive method of interest, their average age, their labour force participation rate and their ethnic distribution in the population. By 'know' here we mean either the actual population value or at least an accurate estimate. The question here is how to efficiently integrate this auxiliary population information into the model fitting process described above.

In some cases, this information is incorporated in the survey weights, through the process of calibration (Deville and Särndal, 1992; Chambers, 1996). That is, these weights are constructed so that weighted averages for selected variables measured in the survey equal corresponding known (or highly accurate estimates of) population values. One approach to using this auxiliary information would therefore be to use such calibrated weights in estimation. However, this has two major problems. First, such weights typically lead to increased standard errors compared to unweighted analysis. Second, survey weights are usually calibrated to a fixed and relatively small set of variables (e.g. age by sex population distributions, regional population distributions), while population data are often known for many more variables.

Alternative, more model-based, ways of incorporating auxiliary population information when modelling survey data have been explored in the econometrics literature, mainly in the context of analysis of linked data sets. An early example is Imbens and Lancaster (1994), who suggest a generalised method of moments approach to the problem of incorporating knowledge of the population expected value of the response variable Y into a sample-based linear regression of Y on an explanatory variable X. More recently, Qin (2000) has considered the same problem using a combination of empirical and parametric likelihood. See Handcock, Rendall and Cheadle (2005) for a comprehensive review of recent developments in this area.

This paper focuses on developing methods for efficiently using auxiliary population information when survey data are used to fit a linear logistic regression model to a target population. In particular, we look at how maximum likelihood methods for fitting such models can be modified to incorporate this information. The approach we take is based on the general approach to maximum likelihood estimation for complex surveys described in Breckling *et. al.* (1994), hereafter referred to as BCDTW. In the following section we develop the general theory for maximum likelihood estimation of a linear logistic model when the sample data are obtained via simple random sampling. In Section 3 we develop corresponding results when case-control sampling is used. In both sections we present simulation results that illustrate the efficiency gains from incorporating auxiliary population information. Section 4 concludes the paper with a discussion of related issues and extensions of the method.

2. MLE under random sampling

Consider the following simple situation. A survey measures the values y_i and x_i of a zero-one variable, Y, and a scalar variable, X, respectively, for a sample s of n units from a population U of N units. The variable X is a population covariate, i.e. we know the values of X

for every unit in the population and the sampling method is simple random sampling without replacement. Our aim is to use these sample data to fit a linear logistic model to the population values of *Y* and *X*. That is, we want to use these data to estimate the parameter $\beta = (\beta_0, \beta_1)^T$ that characterises the linear logistic population regression model

$$\pi(x_i) = \Pr(y_i = 1 \mid x_i) = \exp(\beta_0 + \beta_1 x_i)\left\{1 + \exp(\beta_0 + \beta_1 x_i)\right\}^{-1}.$$

Assuming distinct population units are independently distributed, the maximum likelihood estimate (MLE) for β is then the solution to the estimating equations

$$sc_{1n}(\beta) = \sum_s \left\{y_i - \pi(x_i)\right\} = 0,$$

$$sc_{2n}(\beta) = \sum_s x_i \left\{y_i - \pi(x_i)\right\} = 0$$

where the summations are over the sampled units. We refer to this estimate as the sample MLE, or SMLE, in what follows to indicate that it is the MLE that only uses the information in the *n* sample values of *Y* and *X*.

Suppose that we also know the population total $t_y = \sum_U y_i$ of *Y*. This can happen, for example, if the variable *Y* is also measured in a census, and census tabulations are published. Suppose as well that the individual population values of *X* are known and identifiable (as would be the case if they were held on a register and the sample selected from that register). Now the SMLE is no longer the MLE for β. In order to obtain the 'full information' MLE that includes this additional information, we first observe that the population level score function for this parameter is $sc_U(\beta) = \left(sc_{U1}(\beta), sc_{U2}(\beta)\right)^T$, where

$$sc_{U1}(\beta) = \sum_U \left\{y_i - \pi(x_i)\right\}$$

$$sc_{U2}(\beta) = \sum_U x_i \left\{y_i - \pi(x_i)\right\}$$

and the summations are over the *N* units defining the population *U*. In what follows we let E_s and Var_s denote the expectation and variance operators respectively that condition on the

'available data' for use in analysis. In this case these data correspond to the sample values of *Y* and *X*, the non-sample values of *X* and the population mean of *Y*. We refer to the score function for β given these data as the *full information* score function for this parameter. BCDTW show that the full information score function is the conditional expectation of the corresponding population level score function given the available data. That is, the full information score function $sc_s(\beta) = \left(sc_{s1}(\beta), sc_{s2}(\beta)\right)^T$, where

$$sc_{s1}(\beta) = E_s\left\{sc_{U1}(\beta)\right\} = \sum_U y_i - \sum_U \pi(x_i) \tag{1a}$$

$$sc_{s2}(\beta) = E_s\left\{sc_{U2}(\beta)\right\} = \sum_s x_i\left\{y_i - \pi(x_i)\right\} + E_s\left(\sum_r x_i y_i\right) - \sum_r x_i \pi(x_i) \tag{1b}$$

where *r* denotes the set of non-sampled population units. Let $t_{ry} = \sum_r y_i = t_y - \sum_s y_i$ denote the known total of *Y* for these non-sample units. Also, for arbitrary non-sample population unit *i*, let *r*(*i*) denote the remaining $N - n - 1$ non-sampled population units. Without loss of generality we assume that $t_{ry} > 0$, and observe that the conditional expectation in (1b) can then be written

$$\begin{aligned} E\left(\sum_r y_i x_i \mid \sum_r y_i = t_{ry}, \mathbf{x}_r\right) &= \sum_r x_i E\left(y_i \mid \sum_r y_j = t_{ry}, \mathbf{x}_r\right) \\ &= \frac{\sum_r x_i \Pr\left(y_i = 1, \sum_{r(i)} y_j = t_{ry} - 1 \mid \mathbf{x}_r\right)}{\Pr\left(\sum_r y_j = t_{ry} \mid \mathbf{x}_r\right)} \\ &= \sum_r x_i \pi(x_i) R_{1i}, \end{aligned}$$

where $R_{1i} = \left\{\Pr\left(\sum_r y_j = t_{ry} \mid \mathbf{x}_r\right)\right\}^{-1} \Pr\left(\sum_{r(i)} y_j = t_{ry} - 1 \mid \mathbf{x}_{r(i)}\right)$. The full information score function components defined by (1) are therefore

$$sc_{s1}(\beta) = \sum_U \left\{y_i - \pi(x_i)\right\} \tag{2a}$$

$$sc_{s2}(\beta) = \sum_s x_i\left\{y_i - \pi(x_i)\right\} - \sum_r x_i \pi(x_i)(1 - R_{1i}). \tag{2b}$$

A saddlepoint approximation to the second term on the right hand side of (2b) is developed in the Appendix. This is

$$sc_{s2}(\beta) \approx \sum_s x_i \{y_i - \pi(x_i)\} - \sum_r x_i \pi(x_i)\left(1 - [1 + \{1 - \pi(x_i)\}\{b(t_{ry}) - 1\}]^{-1}\right) \qquad (2c)$$

with $b(t_{ry}) = \exp\left(\left[\sum_r \pi(x_j)\{1 - \pi(x_j)\}\right]^{-1}\{\sum_r \pi(x_j) - t_{ry}\}\right)$. We denote by FIMLE the corresponding approximation to the full information MLE of β obtained by setting (2a) and (2c) to zero and solving for β.

It is unlikely in practice that the actual non-sample *X* values will be known. Since the full information score function (1) depends directly on these values, we need to revise this function when non-sample *X* values are unavailable. In general, the score function for β is then defined by

$$sc_{s1}(\beta) = \sum_U y_i - \sum_s \pi(x_i) - E_s\left\{\sum_r \pi(x_i)\right\} \qquad (3a)$$

$$sc_{s2}(\beta) = \sum_s x_i \{y_i - \pi(x_i)\} + E_s\left(\sum_r x_i y_i\right) - E_s\left\{\sum_r x_i \pi(x_i)\right\} \qquad (3b)$$

where E_s now denotes expectation after conditioning on the actual auxiliary information that we have (we continue to assume that t_{ry} is known). Suppose that we know the non-sample mean $\bar{x}_r$ of *X*. We can then approximate the conditional expectations $E_s\left\{\sum_r \pi(x_i)\right\}$ and $E_s\left\{\sum_r x_i \pi(x_i)\right\}$ in (3) using a smearing approach (Duan, 1983). This is based on the assumption that, for an arbitrary function f of *x* that depends on some parameter θ, we can write

$$\frac{1}{N-n}\sum_r f(x_i, \theta) = \frac{1}{N-n}\sum_r f\left(\bar{x}_r + (x_i - \bar{x}_r), \theta\right) \approx \frac{1}{n}\sum_s f\left(\bar{x}_r - \bar{x}_s + x_i, \theta\right).$$

Put $\Delta = \bar{x}_r - \bar{x}_s$. The smearing approximation to $E_s\left\{\sum_r \pi(x_i)\right\}$ is then

$$E_s\left\{\sum_r \pi(x_i)\right\} \approx \frac{N-n}{n}\sum_s \pi(\Delta + x_i).$$

We therefore replace the score component (3a) by

$$sc_{smear1}(\beta) \approx \sum_U y_i - \sum_s \pi(x_i) - \frac{N-n}{n}\sum_s \pi(\Delta + x_i)\,. \tag{4a}$$

A corresponding smearing approximation to (3b) that includes a saddlepoint approximation is given by (A.7) in the Appendix. This allows us to replace this component score by

$$\begin{aligned} sc_{smear2}(\beta) = {} & \sum_s x_i\left(y_i - \pi(x_i)\right) - \left(\frac{N-n}{n}\right)\sum_s \left(\Delta + x_i\right)\pi(\Delta + x_i) \\ & + \left(\frac{N-n}{n}\right)\sum_s \left(\Delta + x_i\right)\pi(\Delta + x_i)\left[1 + \left\{1 - \pi(\Delta + x_i)\right\}\left\{b_{smear}(t_{ry}) - 1\right\}\right]^{-1} \end{aligned} \tag{4b}$$

where

$$b_{smear}(t_{ry}) = \exp\left(\left[\sum_s \pi(\Delta + x_i)\left\{1 - \pi(\Delta + x_i)\right\}\right]^{-1}\left\{\sum_s \pi(\Delta + x_i) - \frac{n}{N-n}t_{ry}\right\}\right).$$

We refer to the solution to setting (4) to zero and solving for β as the SMEAR estimator for this parameter.

In practice population 'benchmarks' like t_y and $\bar{x}_U$ may in fact be estimated. This can arise, for example, if census coverage is incomplete, and so census outputs are adjusted for coverage error. It can also be the case that we have access to estimates derived from another larger survey rather than census values for these benchmarks. As long as the error or imprecision of such estimation is small, the estimator defined by (4) is still valid. In particular, if benchmark estimates $(\tilde{t}_y, \tilde{x}_U)$ for $(t_y, \bar{x}_U)$ that satisfy $\tilde{t}_y = t_y + o_p(n^{1/2})$ and $\tilde{x}_U = \bar{x}_U + o_p(n^{1/2}/N)$ are used in (4), then, apart from a negligible error of $o_p(n^{-1/2})$, the resulting estimate of β is asymptotically equivalent to the SMEAR estimator. This comment also applies to the FIMLE when the benchmark t_y is assumed to be subject to error.

Finally, there is the case where even $\bar{x}_r$ is unknown. In this case we can still use (4), but replace $\bar{x}_r$ by an appropriate sample-based estimate. This will depend on the characteristics of the sample design and the nature of the auxiliary population information available to us. For the case of simple random sampling and no auxiliary information it is

natural to estimate $\bar{x}_r$ by $\bar{x}_s$, i.e. use expansion estimation. This is equivalent to setting $\Delta = 0$ in (4). We refer to the estimator of β obtained by replacing $\bar{x}_r$ by $\bar{x}_s$ and setting (4) to zero as the expansion MLE for this parameter, and denote it by EXP.

Results from a simulation study of the performance of the estimators described above are set out in Table 1. A total of 1,000 independent simulations were carried out, with N = 5,000 population values for X generated from the standard lognormal distribution and corresponding values for Y generated under the linear logistic model. A sample of n = 200 was then taken from each population using simple random sampling without replacement (SRSWOR). The impact of benchmark error on these relative efficiencies was assessed by considering three levels of imprecision in the benchmark values used as auxiliary information – no error in the benchmarks, benchmarks subject to census-level error (benchmark used equal to true value plus a random error with zero mean and standard deviation equal to the actual marginal standard deviation divided by $N^{-1/2}$) and benchmarks subject to with larger survey error (benchmark used equal to true value plus a random error with zero mean and with standard deviation equal to the actual marginal standard deviation divided by $(N/5)^{-1/2}$).

The values shown in Table 1 are relative efficiencies, defined as the ratio of the 5% trimmed root mean squared error (5%RMSE) of a reference estimator to the corresponding 5%RMSE of an alternative estimator, expressed as a percentage. Values over 100 therefore indicate superior relative efficiency for the alternative estimator. The 5%RMSE is the square root of the 5% trimmed mean of the squared errors generated by an estimator, i.e. after trimming the top 5% and bottom 5% of these squared errors. A trimmed RMSE was used to measure efficiency in order to avoid distortions caused by a small number of outlying error values generated in the simulations. The reference estimation method in Table 1 is SMLE (i.e. the usual estimator of β given sampling is SRSWOR), computed using the *glm* function in R

with its default options. The EXP, SMEAR and FIMLE estimators were calculated by using the *nlm* function in R to solve their respective estimating equations, with starting values $\beta_0 = \log(\bar{y}_U) - \log(1 - \bar{y}_U)$ and $\beta_1 = 0$.

Table 1 about here

From Table 1 we clearly see that, even when the benchmark data contain errors, all three estimation methods that use this information are superior to standard logistic modelling in terms of efficiency. Furthermore, there is little to choose between any of the three estimation methods, with the FIMLE method marginally superior at least for estimation of β_0.

3. MLE under case-control sampling

In the previous section we assumed simple random sampling from the population of interest. However, in many important applications of logistic modelling, particularly in medicine, the sample data are obtained via some form of case-control sampling. In this situation the assumptions underpinning the saddlepoint and smearing approximations used in the development in the previous section are no longer valid. However, the basic strategy of using the approach of BCDTW to incorporate auxiliary population information into inference can still be used, provided the fact that the sample data are obtained via an informative sampling method (case-control sampling) is allowed for when taking conditional expectations. More specifically, we adopt the setup described in Scott and Wild (1997), and assume the existence of two sampling frames, one for the N_1 population units with values $Y = 1$ and one for the N_0 units with $Y = 0$. Independent simple random samples of size n_1 and n_0 respectively are then taken from these frames. Values of X are observed on the sample, and the aim again is to fit a linear logistic model to these data. By definition, we know N_1 and hence $t_{ry} = N_1 - n_1$.

Again, we consider the same three situations corresponding to different levels of knowledge of *X*. The first is where we know the non-sample values of this variable. In the standard case-control situation this is highly unlikely. However, it could correspond to a situation where a separate administrative register contains these values, and the case-control study is being used to forge a link between the *Y* registers and the *X* register. The second is where no *X* register exists, but the value of $\bar{x}_r$ (or an accurate estimate of this quantity) is known. The third is the conventional case-control situation, where no *X* knowledge is available outside the sample. In all three cases, the ML estimating equations for the parameter β of the assumed population level linear logistic model are theoretically defined as the conditional expectations of the population level ML estimating equations given the sample data and the known population information. However, in this case the random variables underpinning these conditional expectations no longer follow the same logistic model as in the population, so the approximations to the ML score function derived in the previous section need modification.

To start, consider the first situation described above, where individual *X* values for non-sample population units are known, but the corresponding values of *Y* are not. We continue to use the notation introduced in the previous section. From (1), we see that the key unknown quantity in the score function is $E_s\left(\sum_r x_i y_i\right)$, where now, because of the case-control sampling, the y_i values in the summation no longer follow the assumed population level logistic model. Following Scott and Wild (1997), we use Bayes Theorem to approximate the distribution of these values as $N - n$ independent Bernoulli realisations with

$$\pi_r(x_i) = \Pr\left(y_i = 1 \mid i \in r, x_i\right) = \frac{N_1^{-1}(N_1 - n_1)\pi(x_i)}{N_1^{-1}(N_1 - n_1)\pi(x_i) + N_0^{-1}(N_0 - n_0)\left\{1 - \pi(x_i)\right\}}.$$

With this set up, we can use the same saddlepoint arguments as in the previous section to approximate $E_s\left(\sum_r x_i y_i\right)$, replacing $\pi(x_i)$ in that development by $\pi_r(x_i)$ above. This leads to a 'full information' score function with component (2a) as before, but with (2c) replaced by

$$sc_{s2}(\beta) = \sum_s x_i\left\{y_i - \pi(x_i)\right\} + \sum_r x_i \pi_r(x_i)\left[1 + \left\{1 - \pi_r(x_i)\right\}\left\{b_r(t_{ry}) - 1\right\}\right]^{-1} - \sum_r x_i \pi(x_i), \qquad (5)$$

where $b_r(t_{ry}) = \exp\left(\left[\sum_r \pi_r(x_i)\left\{1 - \pi_r(x_i)\right\}\right]^{-1}\left\{\sum_r \pi_r(x_i) - t_{yr}\right\}\right)$.

In the previous sub-section, we used smearing to approximate the score function in the case where the individual non-sample *X* values are unknown, but their mean $\bar{x}_r$ is known. This approach needs modification under case-control, because sample and non-sample averages no longer have the same expected values. In particular, for the case-control design assumed here, we need to apply smearing approximations separately for cases and controls. That is, for an arbitrary function f of x characterised by a parameter θ, we use the approximation

$$\sum_r f(x_i,\theta) \approx (N_1 - n_1)n_1^{-1}\sum_{s1} f\left(\Delta_1 + x_i,\theta\right) + (N_0 - n_0)n_0^{-1}\sum_{s0} f\left(\Delta_0 + x_i,\theta\right).$$

Here sd denotes the sample units with $Y = d$ and Δ_d denotes our best estimate of the difference between the non-sample and sample means of *X* for those units with $Y = d$. Since we know the overall non-sample mean $\bar{x}_r$ of *X*, we calculate Δ_d using a regression type estimate, i.e.

$$\Delta_d = \lambda_d n_d^{-1} s_{xd}^2\left(\lambda_1^2 n_1^{-1} s_{x1}^2 + \lambda_0^2 n_0^{-1} s_{x0}^2\right)^{-1}\left(\bar{x}_r - \lambda_1 \bar{x}_{s1} - \lambda_0 \bar{x}_{s0}\right)$$

where $\lambda_d = \left(N_d - n_d\right)/(N - n)$ and $\bar{x}_{sd}$, s_{xd}^2 denote the mean and variance of *X* for the sample units with $Y = d$. The case-control version of the smearing approximation (4a) is then

$$sc_{smear1}(\beta) = \sum_U y_i - \sum_s \pi(x_i) - \sum_{d=0}^{1} \frac{N_d - n_d}{n_d}\sum_{sd} \pi(\Delta_d + x_i) \qquad (6a)$$

while the corresponding case-control version of (4b) is

$$sc_{smear2}(\beta) = \sum_s x_i \left(y_i - \pi(x_i) \right) - \sum_{d=0}^{1} \left(\frac{N_d - n_d}{n_d} \right) \sum_{sd} \left(\Delta_d + x_i \right) \pi(\Delta_d + x_i)$$
$$+ \sum_{d=0}^{1} \left(\frac{N_d - n_d}{n_d} \right) \sum_s \left(\Delta_d + x_i \right) \pi_r(\Delta_d + x_i) \left[1 + \left\{ 1 - \pi_r(\Delta_d + x_i) \right\} \left\{ b_{smear}^{cc}(t_{ry}) - 1 \right\} \right]^{-1} \quad (6b)$$

where

$$b_{smear}^{cc}(t_{ry}) = \exp\left(\left[\sum_{d=0}^{1} \frac{N_d - n_d}{n_d} \sum_{sd} \pi_r(\Delta_d + x_i) \left\{ 1 - \pi_r(\Delta_d + x_i) \right\} \right]^{-1} \left\{ \sum_{d=0}^{1} \frac{N_d - n_d}{n_d} \sum_{sd} \pi_r(\Delta_d + x_i) - t_{yr} \right\} \right).$$

When $\bar{x}_r$ is also unknown, we replace $\bar{x}_{rd}$ by $\bar{x}_{sd}$ above. This is equivalent to setting $\Delta_d = 0$ in (20) and corresponds to using stratified expansion estimators for the expected values of the unknown non-sample components of the score function.

In what follows we use the same notation as in the previous section, denoting estimates obtained by setting (2a) and (5) to zero by FIMLE, and referring to them as full information MLEs. Estimates obtained by setting (6) to zero and solving are referred to as smearing MLEs and are denoted by SMEAR. Finally, those obtained by solving (6) with $\Delta_d = 0$ are referred to as expansion MLEs and are denoted by EXP.

Table 2 sets out simulation results for the above estimators as well as for the standard sample-based MLE of β_1 (denoted SMLE). Prentice and Pyke (1979) showed that the SMLE of β_1 provides a good approximation to the actual MLE of this parameter under case-control sampling. We do not provide results for the SMLE of β_0 since, as is well known, this estimator is seriously biased under case-control sampling. The entries in Table 2 are relative 5%RMSEs, where the reference estimation method is maximum pseudo-likelihood, defined by solving weighted versions of the SMLE estimating equations, with weights given by $w_i = N_0 n_0^{-1} I(y_i = 0) + N_1 n_1^{-1} I(y_i = 1)$, and denoted by WTD. We also computed the maximum 'pseudo-model' likelihood estimates proposed by Scott and Wild (1997) for case-control

sampling, but do not show results for them since these were almost identical to those for the SMLE for β_1 and tended to be unstable for β_0.

The simulation methodology used to obtain the results in Table 2 is identical to that used in Table 1, with the exception that sampling here is carried out using the stratified case-control design described at the start of this section. The SMLE and WTD estimates were computed using the *glm* function in R (without and with weights respectively) with default settings. The FIMLE, SMEAR and EXP estimators were all computed by using the *nlm* function in R to solve the relevant estimating equations.

Table 2 about here

The results set out in Table 2 confirm once again that inclusion of population level auxiliary information can bring substantial gains in maximum likelihood-based inference. This is particularly the case where this information is strong, as in the FIMLE. However, there are still substantial gains when the auxiliary information used is much weaker, as in SMEAR and EXP.

4. Discussion

The two most important conclusions that we draw from the results set out in this paper is that it pays to include population level auxiliary information when modelling sample survey data, and that the BCDTW likelihood framework offers a viable approach to achieving this aim. Obviously, the more auxiliary information one has available, the more significant the improvement in one's inference. However, even marginal information (e.g. knowledge of population means for the model variables) can be very useful when integrated with the sample data within this framework.

An important aspect of auxiliary information is its accuracy. In this paper we use simulation to assess the sensitivity of likelihood methods to error in population benchmarks.

This is because there is usually little or no knowledge of the mechanism underlying benchmark error, and so one has to accept benchmarks 'on trust'. In this context we note that our likelihood methods appear generally quite robust to benchmark error. However, if we do have information about the process that gave rise to these errors, then this information can to be included in the conditioning process underpinning the BCDTW framework. For example, if the observed population means of Y and X are their actual population values plus normally distributed errors with zero means and known variances, then we can modify the development leading to the FIMLE and SMEAR estimators to condition on these observed means, rather than the true ones. This leads to different sets of estimating equations for β that depend on these known variances. A much more difficult problem is where the benchmark errors are 'true' population values, but are biased, e.g. because of subtle differences in the way Y and X are measured in the survey and in the source of the benchmarks. Further research is necessary to see whether this type of auxiliary population information is useful in inference.

In general, use of the BCDTW framework requires one to evaluate conditional expectations that depend both on the assumed population model as well as on the method used to select the sample. For the important case of a logistic population model, the saddlepoint and smearing approximations to these conditional expectations that we describe in this paper seem to work well and should be useful in extending our results in practice. While the results in this paper have been presented in the framework where X is a scalar for simplicity, it is straightforward to put them in the multiple regression settings. Extension of our approach to more complex sampling designs (e.g. unequal probability sampling of controls in case-control situations) should be possible, but will require appropriate adjustments to the saddlepoint and smearing approximations that we use.

This paper does not include results on interval estimation when auxiliary population data are integrated into likelihood inference. The BCDTW framework also covers this

situation, and in the Appendix we show how the information function can be extended to allow for the auxiliary information in the case of a logistic model, including appropriate saddlepoint approximations. An important use of this function is in evaluating the extra information for parametric inference provided by the auxiliary information, e.g. along the lines set out in Steel *et. al.* (2004).

Finally, we point out that our use of the BCDTW framework for integration of auxiliary population information into maximum likelihood inference is quite general. We have focussed on the logistic model situation in this paper because of its practical applications. However, other types of modelling, e.g. linear modelling, can benefit just as much from use of this auxiliary information. In fact, we have obtained parallel results for the linear model case that support this argument. Details are available from the authors on request.

Acknowledgement

Part of this research was carried out when Wang was at the University of Wollongong as a Visiting Professorial Fellow. In addition, this research was supported in part by the TAMU Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30 ES09106).

Appendix

A. Saddlepoint Approximations

We first consider an approximation of R_{1i}. Let $\bar{y}_v$ be the mean of Y over the set v, with N_v the corresponding number of observations. Further, let $g_v(d) = \Pr(\bar{y}_v = d \mid \mathbf{x}_v)$ and $\pi_i = \pi(x_i)$. Then, for $t_{ry} > 0$

$$R_{1i} = \frac{g_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\}}{\pi_i g_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\} + (1-\pi_i) g_{r(i)}(t_{ry}/N_{r(i)})} = \left(1 + (1-\pi_i)\left[\frac{g_{r(i)}(t_{ry}/N_{r(i)})}{g_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\}} - 1\right]\right)^{-1}.$$

(A.1)

It follows that the major problem is to approximate $\left[g_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\}\right]^{-1} g_{r(i)}(t_{ry}/N_{r(i)})$ accurately. The cumulant generating function of $\sum_v y_j$ is $K_v(u) = \sum_v \log\{\pi_j e^u + (1-\pi_j)\}$. For any $d \in (0,1)$ the saddlepoint approximation to $g_v(d)$ is then

$$h_v(d) = \frac{N_v}{\{2\pi K_v''(u_d)\}^{1/2}} \exp\{K_v(u_d) - N_v u_d d\},$$

where u_d is called the saddlepoint, and is defined as the solution of

$$K_v'(u)/N_v = d. \qquad \text{(A.2)}$$

Standard arguments can be used to show that $h_v(d) = g_v(d)\{1 + O(\frac{1}{N_v})\}$ under general regularity conditions. That is, the saddlepoint approximation has a relative error of order N_v^{-1}. Substituting $d = d_1 = t_{ry}/N_{r(i)}$ or $d = d_2 = (t_{ry}-1)/N_{r(i)}$ in $h_{r(i)}(d)$, we then have

$$\frac{g_{r(i)}(t_{ry}/N_{r(i)})}{g_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\}} = \frac{h_{r(i)}(t_{ry}/N_{r(i)})}{h_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\}}\{1 + O(\tfrac{1}{N})\} = \exp\{-u_{d_1}\}\{1 + O(\tfrac{1}{N})\}, \qquad \text{(A.3)}$$

where the last equation is due to the identity

$$K_{r(i)}(u_{d_1}) - N_{r(i)} u_{d_1} d_1 - \left\{K_{r(i)}(u_{d_2}) - N_{r(i)} u_{d_2} d_2\right\} = N_{r(i)} u_{d_1}(d_2 - d_1) + O(\tfrac{1}{N}) = -u_{d_1} + O(\tfrac{1}{N}).$$

From the central limit theorem $N_v^{-1/2}\sum_v (y_j - \pi_j) \to N(0,\gamma^2)$ as $N_v \to \infty$, where $\gamma^2 = \lim N_v^{-1}\sum_v \pi_j(1-\pi_j)$. It follows that we can focus on the normal deviation values of t_{ry}: $t_{ry} - \sum_{r(i)} \pi_j = O\left(\sqrt{N}\right)$. For such values of t_{ry}, $u_{d_1} = O\left(N^{-1/2}\right)$. In fact, from (A.2),

$$u_{d_1} = \frac{t_{ry} - \Sigma_{r(i)}\pi_j}{\Sigma_{r(i)}\pi_j(1-\pi_j)} + O\left(\frac{1}{N}\right) = \frac{t_{ry} - \Sigma_r \pi_j}{\Sigma_r \pi_j(1-\pi_j)} + O\left(\frac{1}{N}\right). \qquad \text{(A.4)}$$

By (A.1), (A.3) and (A.4), an approximation to R_{1i} is then

$$R_{1i} = \left[1 + (1-\pi_i)\left\{b(t_{ry}) - 1\right\}\right]^{-1}\left\{1 + O\left(\frac{1}{N}\right)\right\} \tag{A.5}$$

with $b(t_{ry}) = \exp\left[\left\{\sum_r \pi_j(1-\pi_j)\right\}^{-1}\left(\sum_r \pi_j - t_{ry}\right)\right]$. It immediately follows that (2b) can be approximated by

$$sc_{s2}(\beta) \approx \sum_s x_i(y_i - \pi_i) - \sum_r x_i\pi_i\left(1 - [1 + (1-\pi_i)\{b(t_{ry}) - 1\}]^{-1}\right). \tag{A.6}$$

When non-sample values of X are unavailable, but their mean $\bar{x}_r$ is known, we can combine the saddlepoint approximation developed above with a smearing approximation to again approximate the logistic score function. In particular, this procedure can be used together with (A.6) to approximate the second part of (3b). We continue to use (4a) to approximate (3a). By (A.6),

$$\begin{aligned} sc_{s2}(\beta) &\approx \sum_s x_i(y_i - \pi_i) - \sum_r \left\{\bar{x}_r + (x_i - \bar{x}_r)\right\}\pi_i\left(1 - \left[1 + (1-\pi_i)\left\{b(t_{ry}) - 1\right\}\right]^{-1}\right) \\ &\approx \sum_s x_i(y_i - \pi_i) - \left(\frac{N-n}{n}\right)\sum_s\left(\bar{x}_r - \bar{x}_s + x_i\right)\pi_{i,adj}\left(1 - \left[1 + (1-\pi_{i,adj})\left\{b(t_{ry}) - 1\right\}\right]^{-1}\right) \\ &\approx \sum_s x_i(y_i - \pi_i) - \left(\frac{N-n}{n}\right)\sum_s\left(\bar{x}_r - \bar{x}_s + x_i\right)\pi_{i,adj}\left(1 - \left[1 + (1-\pi_{i,adj})\left\{b_{adj}(t_{ry}) - 1\right\}\right]^{-1}\right) \end{aligned} \tag{A.7}$$

where

$$\pi_{i,adj} = \exp\left\{\beta_1(\bar{x}_r - \bar{x}_s) + \beta_0 + \beta_1 x_i\right\} / \left[1 + \exp\left\{\beta_1(\bar{x}_r - \bar{x}_s) + \beta_0 + \beta_1 x_i\right\}\right]$$

and

$$b_{adj}(t_{ry}) = \exp\left[\left\{\sum_s \pi_{i,adj}\left(1 - \pi_{i,adj}\right)\right\}^{-1}\left(\sum_s \pi_{i,adj} - \frac{n}{N-n}t_{ry}\right)\right].$$

Note that the last two approximation steps in (A.7) used smearing approximations repeatedly.

B. The Information Function

Within the BCDTW framework the information function for parametric likelihood inference is the conditional expectation of the population level information function,

$info_U(\beta)$, minus the conditional variance of the corresponding population level score function. As always, this conditioning is with respect to the observed survey data as well as the auxiliary information. For the simple logistic model considered in this paper the components of the population information function are defined by the decomposition

$$info_U(\beta) = \begin{bmatrix} info_{U11}(\beta) & info_{U12}(\beta) \\ info_{U12}(\beta) & info_{U22}(\beta) \end{bmatrix}.$$

We use a subscript of *s* to denote the corresponding components of the information function defined by our available data. These are

$$\begin{aligned} info_{s11}(\beta) &= E_s\left\{info_{U11}(\beta)\right\} - Var_s\left\{sc_{U1}(\beta)\right\} \\ &= E_s\sum\nolimits_U \pi(x_i)\left\{1-\pi(x_i)\right\} - Var_s\left[\sum\nolimits_U \left\{y_i - \pi(x_i)\right\}\right] \\ &= \sum\nolimits_U \pi(x_i)\left\{1-\pi(x_i)\right\}, \end{aligned}$$

$$\begin{aligned} info_{s12}(\beta) &= E_s\left\{info_{U12}(\beta)\right\} - Cov_s\left\{sc_{U1}(\beta), sc_{U2}(\beta)\right\} \\ &= E_s\sum\nolimits_U x_i\pi(x_i)\left\{1-\pi(x_i)\right\} - Cov_s\left[\sum\nolimits_U \left\{y_i - \pi(x_i)\right\}, \sum\nolimits_U x_i\left\{y_i - \pi(x_i)\right\}\right] \\ &= \sum\nolimits_U x_i\pi(x_i)\left\{1-\pi(x_i)\right\}, \end{aligned}$$

$$\begin{aligned} info_{s22}(\beta) &= E_s\left\{info_{U22}(\beta)\right\} - Var_s\left\{sc_{U2}(\beta)\right\} \\ &= \sum\nolimits_U x_i^2\pi(x_i)\left\{1-\pi(x_i)\right\} - Var_s\left[\sum\nolimits_U x_i\left\{y_i - \pi(x_i)\right\}\right] \\ &= \sum\nolimits_U x_i^2\pi(x_i)\left\{1-\pi(x_i)\right\} - Var_s\left(\sum\nolimits_U x_i y_i\right), \end{aligned}$$

where

$$\begin{aligned} Var_s\left(\sum\nolimits_U y_i x_i\right) &= Var\left(\sum\nolimits_r y_i x_i \mid \sum\nolimits_r y_i = t_{ry}, \mathbf{x}_r\right) \\ &= E\left(\sum\nolimits_{i\in r}\sum\nolimits_{j\in r} y_i y_j x_i x_j \mid \sum\nolimits_r y_i = t_{ry}, \mathbf{x}_r\right) - \left\{E\left(\sum\nolimits_r y_i x_i \mid \sum\nolimits_r y_i = t_{ry}, \mathbf{x}_r\right)\right\}^2 \end{aligned}$$

with

$$\begin{aligned} E\left(\sum\nolimits_{i\in r}\sum\nolimits_{j\in r} y_i y_j x_i x_j \mid \sum\nolimits_r y_i = t_{ry}, \mathbf{x}_r\right) &= \sum\nolimits_r x_i^2 E\left(y_i \mid \sum\nolimits_r y_k = t_{ry}, \mathbf{x}_r\right) \\ &\quad + \sum\nolimits_{i\in r}\sum\nolimits_{j\neq i\in r} x_i x_j E\left(y_i y_j \mid \sum\nolimits_r y_j = t_{ry}, \mathbf{x}_r\right) \\ &= \sum\nolimits_r x_i^2\pi(x_i)R_{1i} + \sum\nolimits_{i\in r}\sum\nolimits_{j\neq i\in r} x_i x_j \pi(x_i)\pi(x_j)R_{2ij} \end{aligned}$$

$$\left\{E\left(\sum_r y_i x_i \mid \sum_r y_i = t_{ry}, \mathbf{x}_r\right)\right\}^2 = \left\{\frac{\sum_r x_i \pi(x_i) \Pr\left(\sum_{r(i)} y_j = t_{ry} - 1 \mid \mathbf{x}_{r(i)}\right)}{\Pr\left(\sum_r y_k = t_{ry} \mid \mathbf{x}_r\right)}\right\}^2$$
$$= \sum_r x_i^2 \pi^2(x_i) R_{1i}^2 + \sum_{i \in r} \sum_{j \neq i \in r} x_i x_j \pi(x_i) \pi(x_j) R_{1i} R_{1j}$$

and

$$R_{2ij} = \left\{\Pr\left(\sum_r y_k = t_{ry} \mid \mathbf{x}_r\right)\right\}^{-1} \Pr\left(\sum_{r(ij)} y_k = t_{ry} - 2 \mid \mathbf{x}_{r(ij)}\right).$$

It follows

$$Var_s\left(\sum_U y_i x_i\right) = \sum_r x_i^2 \pi(x_i) R_{1i} \left\{1 - \pi(x_i) R_{1i}\right\} + \sum_{i \in r} \sum_{j \neq i \in r} x_i x_j \pi(x_i) \pi(x_j) (R_{2ij} - R_{1i} R_{1j}).$$

A saddlepoint approximation to R_{2ij} similar to that developed above for R_{1i} can be written down. This is based on the fact that the denominator of R_{2ij} can be expressed as

$$\Pr\left(\sum_r y_k = t_{ry} \mid \boldsymbol{x}_r\right) = \pi_i \pi_j \Pr\left(\sum_{r(ij)} y_k = t_{ry} - 2 \mid \boldsymbol{x}_{r(ij)}\right)$$
$$+ \left\{\pi_i (1 - \pi_j) + (1 - \pi_i) \pi_j\right\} \Pr\left(\sum_{r(ij)} y_k = t_{ry} - 1 \mid \boldsymbol{x}_{r(ij)}\right)$$
$$+ (1 - \pi_i)(1 - \pi_j) \Pr\left(\sum_{r(ij)} y_k = t_{ry} \mid \boldsymbol{x}_{r(ij)}\right)$$

leading to

$$R_{2ij} = \left\{\begin{array}{c} \pi_i \pi_j + (\pi_i + \pi_j - 2\pi_i \pi_j) \dfrac{\Pr\left(\Sigma_{r(ij)} y_k = t_{ry} - 1 \mid \boldsymbol{x}_{r(ij)}\right)}{\Pr\left(\Sigma_{r(ij)} y_k = t_{ry} - 2 \mid \boldsymbol{x}_{r(ij)}\right)} \\ + (1 - \pi_i)(1 - \pi_j) \dfrac{\Pr\left(\Sigma_{r(ij)} y_k = t_{ry} \mid \boldsymbol{x}_{r(ij)}\right)}{\Pr\left(\Sigma_{r(ij)} y_k = t_{ry} - 2 \mid \boldsymbol{x}_{r(ij)}\right)} \end{array}\right\}^{-1}.$$

Using the same saddlepoint approximation technique as that used for R_{1i}, the two ratios in this expression can be approximated by $b(t_{ry} - 1)$ and $b^2(t_{ry} - 1)$ respectively. That is,

$$R_{2ij} = \left\{\pi_i \pi_j + (\pi_i + \pi_j - 2\pi_i \pi_j) b(t_{ry} - 1) + (1 - \pi_i)(1 - \pi_j) b^2(t_{ry} - 1)\right\}\left\{1 + O\left(\tfrac{1}{N}\right)\right\}.$$

References

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M. and Welsh, A.H. (1994). Maximum likelihood inference from survey data. *International Statistical Review*, **62**, 349 - 363.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, **12**, 3 - 32.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376 - 382.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation estimate. *Journal of the American Statistical Association*, **78**, 605 - 610.

Handcock, M., Rendall, M. and Cheadle, J. (2005). Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. *Sociological Methodology,* **35**, 291 - 334.

Imbens, G.W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies*, **61**, 655 - 680.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403 - 411.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika*, **87**, 484 - 490.

Scott, A.J. and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, **84**, 57 - 71.

Steel, D.G., Beh, E. J. and Chambers, R.L. (2004). The information in aggregate data. In *Ecological Inference: New Methodological Strategies*. (eds. G. King, O. Rosen and M. Tanner). Cambridge University Press: Cambridge.

Table 1 Linear logistic population model under SRSWOR and population benchmarks of varying quality. Values in table are percent relative efficiencies with respect to 5% trimmed root mean squared error of SMLE. Values of X drawn from the standard lognormal distribution. In all cases N = 5000 and n = 200. Values of X drawn from the standard lognormal distribution.

True (β_0 , β_1)		(–3, 1)	(–5, 2)	(–5, 1)	(–8, 2)
		Population Benchmarks Known Precisely			
β_0	EXP	103.80	103.26	109.35	115.44
	SMEAR	111.75	107.61	114.19	115.43
	FIMLE	115.88	111.81	121.26	113.66
β_1	EXP	101.88	105.49	108.08	118.69
	SMEAR	101.67	104.20	106.35	116.27
	FIMLE	100.95	102.21	105.15	110.42
		Population Benchmarks Subject to Census-level Error			
β_0	EXP	111.62	100.12	113.05	122.63
	SMEAR	115.14	106.55	118.15	121.50
	FIMLE	121.12	111.88	120.90	117.38
β_1	EXP	102.70	103.55	107.55	126.47
	SMEAR	101.25	104.29	106.24	122.25
	FIMLE	102.34	103.43	104.11	115.13
		Population Benchmarks Subject to Larger Survey Error			
β_0	EXP	104.89	98.94	107.36	127.49
	SMEAR	109.69	99.82	110.61	125.51
	FIMLE	114.73	109.53	113.50	121.56
β_1	EXP	101.90	105.28	107.48	129.88
	SMEAR	101.94	104.22	106.53	125.36
	FIMLE	101.88	104.91	106.49	119.95

Table 2 Linear logistic model under case-control sampling with population benchmarks known precisely. Values in table are percent relative efficiencies with respect to 5% trimmed root mean squared error of WTD. In all cases $N = 5000$ and $n_1 = n_0 = 100$. Values of X drawn from the standard lognormal distribution.

True (β_0, β_1)		(–3, 1)	(–5, 2)	(–5, 1)	(–8, 2)
β_0	EXP	105.46	107.63	121.18	127.32
	SMEAR	105.78	108.55	120.07	125.93
	FIMLE	107.75	112.15	144.36	161.02
β_1	SMLE	106.13	107.88	126.78	129.19
	EXP	105.81	107.88	117.31	124.65
	SMEAR	112.92	111.55	127.69	126.96
	FIMLE	121.06	123.13	190.82	189.26